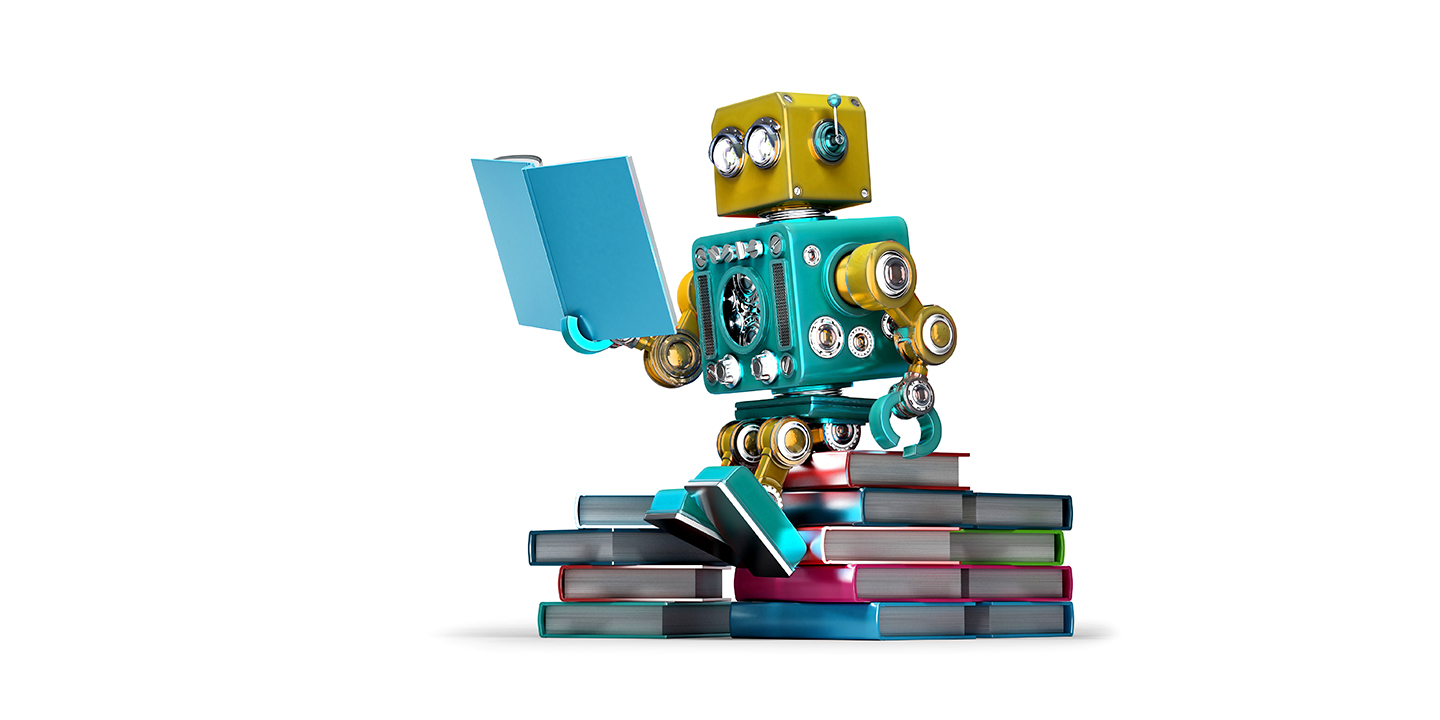
**Μηχανική Μάθηση Εργασία 1**



**Μπόζας Αριστείδης**

**ΑΜ:740**

**Γιάννης Δερμετζόγου**

**ΑΜ:**

**Περιεχόμενα**

[1. Μέρος Α 4](#_Toc511454016)

[2. Μέρος Β 4](#_Toc511454017)

[3. Μέρος Γ 5](#_Toc511454018)

[Παράρτημα κώδικα 1 5](#_Toc511454019)

[Παράρτημα κώδικα 2 5](#_Toc511454020)

[Παράρτημα κώδικα 3 5](#_Toc511454021)

Κατάλογων Πινάκων

Κατάλογων Εικόνων

# Μέρος Α

Στο πρώτο μέρος της εργασίας εξετάστηκε το αντικείμενο των τεχνικών πολλαπλών μοντέλων πρόβλεψης σε συνδυασμό με το αντικείμενο της συγκριτικής αξιολόγησης μεταξύ αλγορίθμων.

Τα 10 datasets που χρησιμοποιήθηκαν από το [UCI](http://archive.ics.uci.edu/ml/index.php) repository είναι τα εξής:

* [Iris](https://archive.ics.uci.edu/ml/datasets/iris)
* [Wine](https://archive.ics.uci.edu/ml/datasets/wine)
* [Wdbc](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))
* [Balance-scale](http://archive.ics.uci.edu/ml/datasets/balance+scale)
* [Hayes roth](https://archive.ics.uci.edu/ml/datasets/Hayes-Roth)
* [Haberman survival](https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival)
* [Liver disorder](https://archive.ics.uci.edu/ml/datasets/liver+disorders)
* [Data bank authedication](https://archive.ics.uci.edu/ml/datasets/banknote+authentication)
* [Ionosphere](https://archive.ics.uci.edu/ml/datasets/ionosphere)
* [Cmc](https://archive.ics.uci.edu/ml/machine-learning-databases/cmc/)

Οι τεχνικές που χρησιμοποιήθηκαν για τα πολλαπλά μοντέλα πρόβλεψη είναι οι εξής:

* **Manipulating the training examples:** Oι τεχνικές του bagging και boosting.
* **Manipulating the target variable:** Oι τεχνικές του OnevsOne και OnevsRest.
* **Injecting randomness :** Tο ensemble μοντέλο RandomForest.
* **Manipulating Features :** Τυχαία επιλογή του των Feature και των παραδειγμάτων εκπαίδευσης με την τεχνική RandomPatches.

Όσο αναφορά το αντικείμενο της συγκριτικής αξιολόγησης μεταξύ των αλγορίθμων που εκτελέστηκαν φαίνονται παρακάτω:

# Μέρος Β

Το δεύτερο μέρος της εργασία εξετάστηκε το πρόβλημα του διαφορετικού κόστους στο σύνολο δεδομένου [heart](http://archive.ics.uci.edu/ml/datasets/statlog+(heart)) .H βιβλιοθήκη που χρησιμοποιήθηκε ήταν η **Costcla,sklearn** της python.

To cost matrix αυτού του συνόλου δεδομένου είναι το εξής:

|  |  |  |
| --- | --- | --- |
|  | **Actual absence** | **Actual presence** |
| **Absence** | 0 | 1 |
| **Presence** | 5 | 0 |

Πίνακας 1: Cost matrix του συνόλου δεδομένου heart

Οι μέθοδοι που χρησιμοποιήθηκαν και αναφέρονται παρακάτω συνδυάστηκαν με τους αλγορίθμους μάθησης Random Forest, Linear SVM,Naive Bayes:

* Δίνοντας έμφαση στα παραδείγματα με το μεγαλύτερο κόστος(**CostSampling** [Oversampling, RejectionSampling], Undersampling )
* Ελαχιστοποίηση αναμενόμενου κόστους εκτιμήσεων (**ThresholdOptimization,BayesMinimumRiskClassifier)**
* Τροποποιημένη cost sensitive ταξινομητές( **CostSensitiveRandoForestClassifier**)

Τα ονόματα στην στήλη Algorithm που φαίνονται στον πίνακα [Χ] αποτελεσμάτων είναι οι τεχνικές που χρησιμοποιήθηκαν σε συντομογραφία πιο συγκεκριμένα παρακάτω δύνονται η πλήρης ονομασίες:

* **RF** : RandomForest
* **RF – O :** RandomForest Over-Sampling
* **RF – R:** RandomForest Rejection-Sampling
* **RF – U:** RandomForest Under-Sampling
* **RF** **- BMR**: RandomForest BayesMinimumRiskClassifier
* **RF - TO**: RandomForest ThresholdOptimization
* **RFC**:CostSensitiveRandoForestClassifier
* **LSVM**: LinearSVM
* **LSVM - O** : LinearSVM Over-Sampling
* **LSVM** **- R**: LinearSVM Rejection-Sampling
* **LSVM** : LinearSVM Under-Sampling
* **LSVM - BMR**: LinearSVM BayesMinimumRiskClassifier
* **LSVM - TO**: LinearSVM ThresholdOptimization
* **GNB**: NaiveBayes
* **GNB – O:** NaiveBayes Over-Sampling
* **GNB – R**: NaiveBayes Rejection-Sampling
* **GNB – U**: NaiveBayes Under-Sampling
* **GNB - BMR**: NaiveBayes BayesMinimumRiskClassifier
* **GNB - TO**:: NaiveBayes ThresholdOptimization

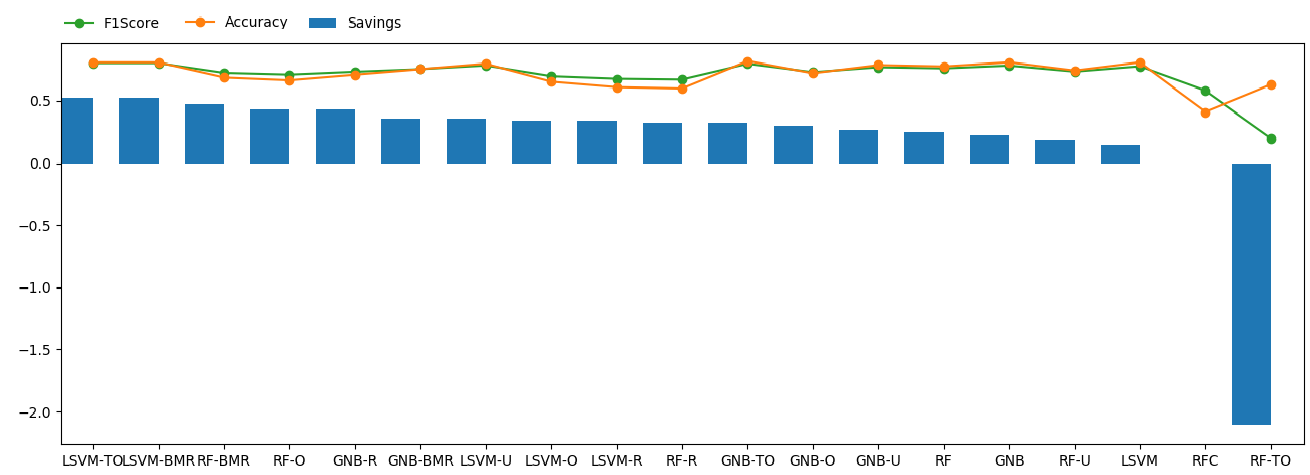
Το SVM τείνει να πιέζει τις προβλεπόμενες πιθανότητες μακριά από 0 και 1. Άλλα μοντέλα όπως Naive bayes έχουν την αντίθετη προκατάληψη και τείνουν να ωθούν τις προβλέψεις πιο κοντά στα 0 και 1.Αυτό βλάπτει την ποιότητα των πιθανοτήτων. Για αυτό έγινε calibration στις πιθανότητες των αλγορίθμων μάθησης Random Forest, Linear SVM,Naive Bayes με τη χρήση της βιβλιοθήκης **sklearn.calibration.**

## 2.1 Αποτελέσματα

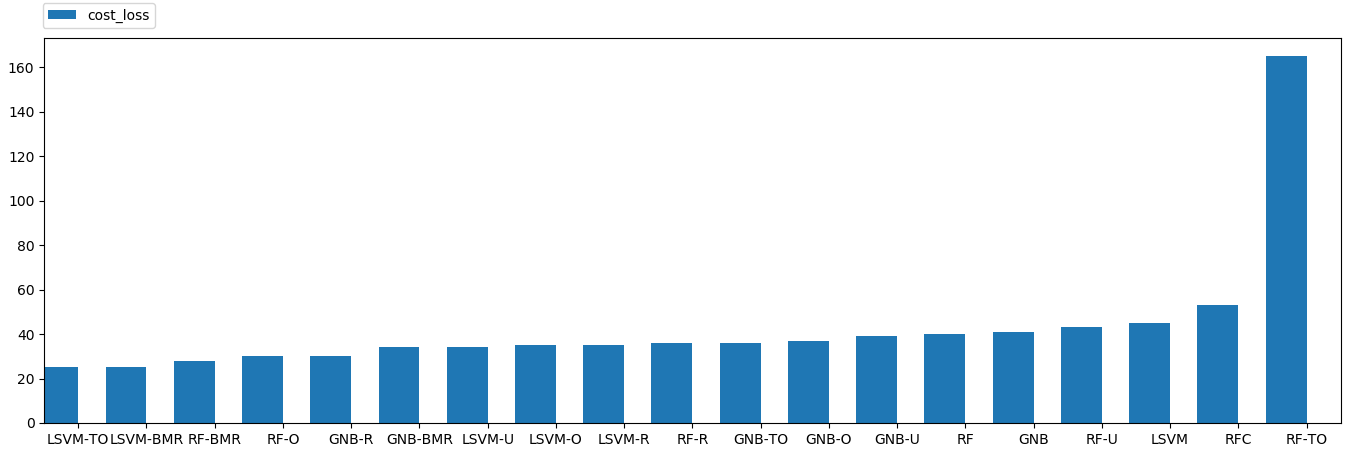
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **F1** | **Accuracy** | **Saving score** | **Cost** |
| **RF** | 0.761905 | 0.777778 | 0.245283 | 40.0 |
| **RF - O** | 0.711538 | 0.666667 | 0.433962 | 30.0 |
| **RF - R** | 0.672727 | 0.600000 | 0.320755 | 36.0 |
| **RF - U** | 0.735632 | 0.744444 | 0.188679 | 43.0 |
| **RF - BMR** | 0.725490 | 0.688889 | 0.471698 | 28.0 |
| **RF - TO** | 0.195122 | 0.633333 | -2.113208 | 165.0 |
| **LSVM** | 0.779221 | 0.811111 | 0.150943 | 45.0 |
| **LSVM - O** | 0.699029 | 0.655556 | 0.339623 | 35.0 |
| **LSVM - R** | 0.678899 | 0.611111 | 0.339623 | 35.0 |
| **LSVM - U** | 0.785714 | 0.800000 | 0.358491 | 34.0 |
| **LSVM - BMR** | 0.804598 | 0.811111 | 0.528302 | 25.0 |
| **LSVM - TO** | 0.804598 | 0.811111 | 0.528302 | 25.0 |
| **GNB** | 0.784810 | 0.738095 | 0.226415 | 41.0 |
| **GNB - O** | 0.731183 | 0.722222 | 0.301887 | 37.0 |
| **GNB – R** | 0.734694 | 0.711111 | 0.433962 | 30.0 |
| **GNB - O** | 0.771084 | 0.788889 | 0.264151 | 39.0 |
| **GNB - BMR** | 0.755556 | 0.755556 | 0.358491 | 34.0 |
| **GNB - TO** | 0.800000 | 0.822222 | 0.320755 | 36.0 |
| **RFC** | 0.582677 | 0.411111 | 0.000000 | 53.0 |

Πίνακας 2: κόστους

Πίνακας 2: Αποτελεσμάτα των μεθόδων για το πρόβλημα του διαφορετικού κόστους



Εικόνα 1:Διάγραμμα saving score προς f1 και accuracy από την καλύτερη απόδοση κόστους προς την χειρότερη



Εικόνα 2:Διάγραμμα κόστους από την καλύτερη απόδοση στη χειρότερη

## Συμπεράσματα

Από τις εικόνες τον αποτελεσμάτων φαίνεται πως το μοντέλο **LSVM – To** και **LVSM - BMR** καθώς έχουν το μικρότερο κόστος και έχουν πολύ καλή απόδοση F1.Αντίθετα το μοντέλο **RF-TO** δίνει τα χειρότερα αποτελέσματα καθώς έχει πολύ μεγάλο κόστος και την χαμηλότερη απόδοση F1.Γενικά, όλα τα μοντέλα ελαχιστοποιήσεις κόστους εκτός το μοντέλο **RF-TO** της δίνουν πολύ καλή απόδοση στη μετρική του κόστους αλλά και στις μετρικής F1.

Ενδιαφέρον παρουσιάζει, ότι δίνοντας έμφαση στα παραδείγματα με το μεγαλύτερο κόστος με τις τεχνικές **Oversampling, RejectionSampling, Undersampling** καθώς χάνουν απόδοση στις μετρικές F1, Αccuracy η μετρική κόστους τους μειώνεται σε σχέση με τους αρχικού αλγορίθμους RF,LSVM,GNB.

Επίσης, οι τροποποιημένη cost sensitive ταξινομητές στην περίπτωση χρησιμοποιήθηκε ο **Cost Sensitive RandoForest** ταξινομητής ο όποιος δεν παρουσιάζει καλά αποτελέσματα στο συγκεκριμένο πρόβλημα.

Τέλος, εκτός της μέθοδος ελαχιστοποίησης του αναμενόμενου κόστους εκτιμήσεων που φαίνεται να διακρίνεται το μοντέλο LSVM.Στις άλλες μεθόδους φαίνεται τα μοντέλα LSVM,GNB να έχουν σχετικά όμοιες αποδώσεις.

# Μέρος Γ

# Παράρτημα κώδικα 1

# Παράρτημα κώδικα 2

# Παράρτημα κώδικα 3